AFHRL-TP-87-70

AIR FORCE HUMAN RESOURCES LABORATORY

GENERALIZABILITY THEORY: AN ASSESSMENT OF ITS RELEVANCE TO THE AIR FORCE JOB PERFORMANCE MEASUREMENT PROJECT

Kurt Kraiger

University of Colorado at Denver
Department of Psychology
1200 Larimer Street
Denver, Colorado 80204

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

April 1989

Interim Technical Paper for Period May 1986 - October 1987

Approved for public release; distribution is unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

HENDRICK W. RUCK, Technical Advisor
Training Systems Division

RODGER D. BALLENTINE, Lt Col, USAF
Chief, Training Systems Division

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release; distribution is unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFHRL-TP-87-70

6a. NAME OF PERFORMING ORGANIZATION
Universal Energy Systems

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Training Systems Division

6c. ADDRESS *(City, State, and ZIP Code)*
4401 Dayton-Xenia Road
Dayton, Ohio 45432

7b. ADDRESS *(City, State, and ZIP Code)*
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
Air Force Human Resources Laboratory

8b. OFFICE SYMBOL *(If applicable)*
HQ AFHRL

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
F41689-84-D-0002

8c. ADDRESS *(City, State, and ZIP Code)*
Brooks Air Force Base, Texas 78235-5601

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO.
62703F	7719	18	40

11. TITLE *(Include Security Classification)*
Generalizability Theory: An Assessment of Its Relevance to the Air Force Job Performance Measurement Project

12. PERSONAL AUTHOR(S)
Kraiger, K.

13a. TYPE OF REPORT
Interim

13b. TIME COVERED
FROM May 86 TO Oct 87

14. DATE OF REPORT *(Year, Month, Day)*
April 1989

15. PAGE COUNT
34

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
12	03	
05	08	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
D-study
G-study
generalizability theory
job performance measurement
reliability

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

Generalizability theory is a method for estimating the dependability of scores over various conditions of measurement. In contrast to classical test theory, which permits the investigation of only one error source at a time, generalizability theory is multifaceted and allows the researcher to propose and simultaneously investigate multiple sources of error. The applicability of generalizability theory to the Air Force's ongoing Job Performance Measurement (JPM) Project is reviewed in this paper. It was concluded that generalizability theory would be relevant to the JPM Project because it is a sophisticated and efficient method of investigating reliability, a fundamental property of any measurement instrument. Next, generalizability theory is illustrated by applying it to data collected from Air Force jet engine mechanics. The question of interest was whether performance scores were generalizable (consistent) over different rating sources (incumbents, supervisors, or peers), rating forms (one of four forms used in the project), or specific items included on any one form. The results indicated that scores were generalizable over both forms and items within forms. However, scores were not generalizable over rating sources. Sources tended to differentially rank ratees, depending on the specific form

Orig

(Continued)

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Nancy J. Allin, Chief, STINFO Branch

22b. TELEPHONE *(Include Area Code)*
(512) 536-3877

22c. OFFICE SYMBOL
AFHRL/SCV

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

Item 19 (Concluded):

used. When scores were averaged over rating sources, the resulting mean scores were generalized if at least two rating forms were used. Implications of the results and the applicability of generalizability theory for the Job Performance Measurement Project are discussed.

AFHRL Technical Paper 87-70

April 1989

GENERALIZABILITY THEORY: AN ASSESSMENT OF ITS RELEVANCE TO THE AIR FORCE JOB PERFORMANCE MEASUREMENT PROJECT

Kurt Kraiger

University of Colorado at Denver
Department of Psychology
1200 Larimer Street
Denver, Colorado 80204

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed and submitted for publication by

Nestor K. Ovalle, II, Lt Col, USAF
Chief, Training Assessment Branch

This publication is primarily a working paper. It is published solely to document work performed.

SUMMARY

This paper reviews generalizability theory and discusses its applicability to the Air Force Job Performance Measurement Project. It is concluded that generalizability theory has relevance for the project. Application of the theory is illustrated with analyses of performance data collected from Air Force jet engine mechanics. Error variance in measurement is estimated for rating sources (incumbents, supervisors, and peers), rating forms, and specific items included on each form. Possible measurement conditions eliciting desirable degrees of generalizability are presented. The relationship of generalizability theory to construct validity and the logical requirements for performance ratings are discussed.

PREFACE

The Air Force Job Performance Measurement Project is a remarkably broad, encompassing attempt to assess individual job proficiency. Within the tested specialties, incumbents are assessed via a Walk-Through Performance Test (with hands-on and interview components), and subjective performance evaluations. The performance ratings themselves are broad, as incumbents are evaluated on four different forms (varying in task specificity) by themselves, their peers, and supervisors.

Collecting so much data for each Air Force specialty would be impractical in the long run. Thus, one goal of the project staff is to reduce the total number of measures collected, by empirically evaluating and comparing the measures as data are collected from new specialties. Another goal is to modify and improve measures which will ultimately be retained. Generalizability theory is ideally suited to both goals. Generalizability theory is useful for investigating whether scores on any measurement instrument are dependable over varying conditions of measurement. Both rating forms and rating sources can be considered measurement conditions. If scores are found to be generalizable, then the number of conditions sampled can be reduced, with minimal losses in generalizability. In fact, generalizability theory can forecast the resulting dependability indices for differing sets of measurement conditions. Thus, it can provide decision-makers with answers to questions such as: How dependable would our rating system be if we used only supervisory ratings on a single form? Similarly, another relevant question would be how the measurement process could be improved by increasing the number of items on each form. Generalizability theory can forecast the effects of dependability for these modifications as well. Thus, generalizability theory appears to have considerable relevancy to the performance measurement project. These issues are discussed and illustrated in this paper.

Several individuals warrant acknowledgement for their assistance in the preparation of this paper. Both Lt Colonel Rodger Ballentine and Dr Jerry Hedge have been extremely helpful during my association with the Laboratory. They have helped pull the strings that were needed on various occasions during the project. Dr Terry Dickinson provided numerous technical comments and useful suggestions. Finally, Mark Teachout has excelled as a contract monitor, pulling the statement of work together, providing data or information upon request, reviewing drafts, and keeping the project close to schedule.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GENERALIZABILITY THEORY: AN ASSESSMENT OF ITS RELEVANCE TO THE AIR FORCE JOB PERFORMANCE MEASUREMENT PROJECT

INTRODUCTION

Generalizability theory was developed by Cronbach and associates (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963) as an alternative to classical test theory. Whereas classical test theory posited a single true score and a single undifferentiated error term, Cronbach and associates recognized that measurement error may be introduced by any number of sources, and advocated multifaceted experiments to estimate the error variance due to each source. Further, they replaced the concept of a single true score with a universe score, the value of which could vary by the nature of the error source and the population of generalization. Thus, generalizability theory offers a more complex, yet more realistic portrayal of measurement error. Generalizability theory contributes to measurement theory in other ways as well, by making certain measurement assumptions explicit. For example, the researcher must specify whether the test is to be used for absolute (criterion-referenced) or relative (norm-referenced) decision-making, as the degree of error will vary by purpose. The many contributions of generalizability theory to measurement theory will be explained below. First, the theory itself must be explained in greater detail. To do so, I will first discuss some ambiguities inherent in classical test theory.

Classical test theory proposes that any observed score for a person on an instrument can be decomposed into two parts: true score and error. Since these two components are assumed to be independent, it follows that observed score variance ($\hat{\alpha}^2_x$) can also be decomposed into true score variance ($\hat{\alpha}^2_t$)and error variance ($\hat{\alpha}^2_e$). The reliability of an instrument is then defined as the proportion of observed score variance which is true score variance ($\hat{\alpha}^2_t/\hat{\alpha}^2_x$). In other words, reliability represents freedom from measurement error ($\hat{\alpha}^2_t/\hat{\alpha}^2_x = \hat{\alpha}^2_x/\hat{\alpha}^2_x - \hat{\alpha}^2_e/\hat{\alpha}^2_x$).

In practice, researchers have operationalized reliability in a number of ways. Each strategy identifies a different source of measurement error and estimates a different "type" of reliability. When a test is administered on two occasions, the correlation between those two sets of scores is called test-retest reliability and it estimates error due to variance in measurement conditions over time. The correlation between scores assigned by two raters or judges is called "conspect reliability," and it estimates error variance introduced by differer: scorers. Though classical test theory provides different descriptors of reliability, it still uses only a single term for each type of error implicit in each form of reliability. Relationships among different kinds of measurement error are unclear and (more critically) inestimable. Classical test theory leaves us with a fundamental paradox of a single true score but multiple estimates of true score variance depending on how error variance is defined. Furthermore, since the theory is univariate by nature, it does not easily allow for the estimation of the joint effects of multiple sources of error variance. Thus, while the Spearman-Brown prophecy formula allows us to predict the resulting reliability of a test after the number of items on the test is increased or decreased, it does not allow us to predict the increase in reliability from an increase in both the length of the test and the number of times it was administered.

Multifaceted Approach of Generalizability Theory

In contrast to classical test theory, generalizability theory explicitly recognizes the existence of multiple sources of error variance and provides methods for simultaneously estimating each. It encourages the researcher to explicitly consider conditions which may affect the measurement process. For example, the ratings one individual receives from a group of peers may depend upon which particular peers were selected as raters, which dimensions were chosen as stimuli, the time of day each peer observed the ratee, or the difficulty of the tasks performed by the ratee at the time of observation. To say that ratings depend on these conditions is to say that the ratee's expected score could change if different elements of each factor were sampled.

In generalizability theory, the researcher identifies the factors affecting measurement which are of the greatest interest or importance. Then, the researcher specifies a particular range of levels of each factor for study. In G theory terminology, factors of measurement are called "facets" and levels of the facet are called "conditions." Thus, in a study of the generalizability of ratings of Walk-Through Performance, two important facets might be the actual tasks performed and the test administrators. That is, the score any ratee receives could depend on the difficulty of the tasks which comprise the Walk-Through Test or on any idiosyncrasies of individual administrators/raters.

A generalizability (G) study could be designed to estimate the contribution of the task and administrator facets to total score variance. Ideally, random samples of tasks and administrators (drawn from the population of all possible tasks and administrators) would be identified and tested on a large sample of individuals. The populations of all possible tasks and administrators define the universe of admissible observations, the boundaries of which must be explicitly defined by the researcher. That is, the researcher would specify what constitutes an acceptable task or unacceptable task for any given administration of the Walk-Through.

The G study could be designed as fully crossed or nested. In a fully crossed design, all conditions of one facet are observed in combination with all conditions of the other facet; i.e., all administrators rate all tasks. With a nested design, different conditions of one facet are nested in, or observed within, different conditions of the other facet. For example, if tasks were nested within administrators, each administrator would observe and rate a different set of tasks. In general, fully crossed designs are preferable because they allow for direct estimation of all possible variance components, but nested designs are often used either because they are more efficient or because they reflect real-world situations. (For example, in most G studies of teacher evaluations, students are nested within classes.)

For the present example, let us assume that we have designed a fully crossed G study in which all individuals are examined on 10 different tasks by four different administrators. This design is illustrated in Figure 1. The cube in Figure 1 represents all cells in the design, or all possible combinations of n persons observed on 10 tasks by four administrators. Note that in contrast to traditional analysis of variance (ANOVA) designs, persons (or subjects) are treated as a factor with only a single replicate in each cell.

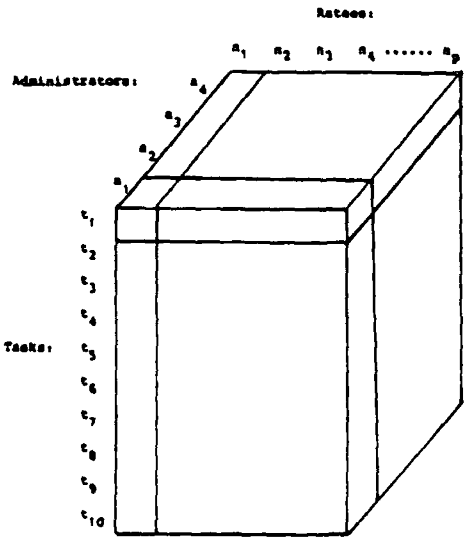

Figure 1. Sample Generalizability Design.

With a set of persons observed in a two-facet fully crossed design, there are seven possible sources of variance. These sources are illustrated by the Venn diagram in Figure 2. In ANOVA terms, there are three "main effects" represented graphically by the variance components for tasks ($\hat{\sigma}^2_t$), administrators ($\hat{\sigma}^2_a$), and ratees or persons ($\hat{\sigma}^2_p$). $\hat{\sigma}^2_t$ and $\hat{\sigma}^2_a$ represent sources of systematic error variance while $\hat{\sigma}^2_p$ represents (desirable) variance due to individual differences. In addition, there are three two-way interaction terms: $\hat{\sigma}^2_{pt}$ and $\hat{\sigma}^2_{pa}$ and $\hat{\sigma}^2_{ta}$. These interactions may be interpreted as follows. Variance due to the task-by-persons interaction ($\hat{\sigma}^2_{pt}$) might be high if persons performed very well on some tasks but very poorly on others. $\hat{\sigma}^2_{ta}$ would indicate error due to differential scoring of tasks by administrators while $\hat{\sigma}^2_{pa}$ would indicate whether persons were differentially ranked by different administrators. The final area of the Venn diagram represents the confounding of the three-way interaction of tasks, administrators, and persons with random error variance ($\hat{\sigma}^2_{pta}+\hat{\sigma}^2_e$). These two sources (variance attributable to the three-way interaction and error variance) cannot be separated since there is only a single observation per "cell" in the design. It is important to note that some of what is called random error variance in this design could be explained as systematic variance in a more complex design with additional facets. Ideally, enough facets could be identified and measured to explain all error variance.

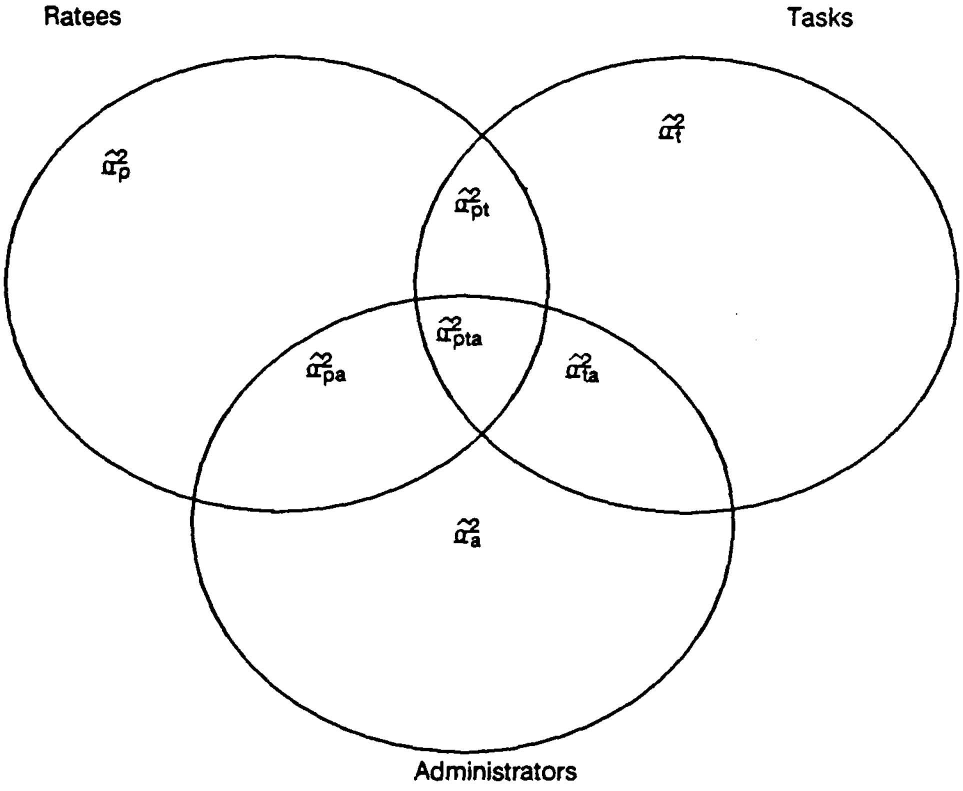

Figure 2. **Venn Diagram Illustrating Variance Components in Sample Design.**

The multifaceted treatment of error variance by generalizability theory should be clearer at this point. For any set of persons, the total observed score variance is represented by the lower full circle in Figure 2. Generalizability theory partitions that variance into individual difference variance ($\hat{\sigma}^2_p$) and multiple, distinct sources of error variance (e.g., $\hat{\sigma}^2_{pt}$ and $\hat{\sigma}^2_{pa}$ and $\hat{\sigma}^2_{pta}$). Just as classical test theory provides a reliability coefficient conceptually equal to $\hat{\sigma}^2_t/\hat{\sigma}^2_x$, generalizability theory provides a generalizability coefficient, $E\rho^2$, which is equal to $\hat{\sigma}^2_p/\hat{\sigma}^2_x$ or $\hat{\sigma}^2_p/(\hat{\sigma}^2_p + \hat{\sigma}^2_{pt} + \hat{\sigma}^2_{pa} + \hat{\sigma}^2_{pta})$ in the present example. However, generalizability coefficients computed from G study estimates are generally not as useful as those generated from a different type of study called a decision (D) study, which is described below. Also, many researchers have stressed that interpretation of the variance components themselves is at least as important as the interpretation of the generalizability coefficient (e.g., Brennan & Kane, 1979).

To return again to the example, we have developed a fully crossed two-facet design which allows estimation of seven distinct sources of variance. After the data are collected, we must compute the actual variance components for each source. Variance components are determined from the mean squares of a traditional ANOVA. In this analysis, persons are treated as a between-subjects factor (with one observation per cell) and the multiple conditions of each facet are treated as repeated measures on a facet. The precise formulas for determining variance component estimates from ANOVA mean squares are often complex, but may be derived from algorithms available in Brennan (1983), Brennan and Kane (1979), or Cronbach et al. (1972).

The variance components computed from a G study represent estimated variance about universe scores for average, or single, observations; e.g., a single person evaluated on a single task by a single administrator. As noted above, these variance components can be used to compare the relative contributions of error variance sources or to compute generalizability coefficients. The results of such analyses may be misleading, however, as the G study design may differ from the way organizations typically use measurement instruments. That is, G study estimates are for single items or administrations, yet organizations typically strive for multiple measures of a construct (or at least multiple-item scales) to minimize error variance relative to universe score variance. Thus, it is necessary to distinguish unitary estimates of G studies from the estimates of decision (D) studies which better reflect how an organization uses a measurement instrument.

Generalizability Analyses for Decision-Making

A G study establishes the general characteristics of a measuring device; in particular, the relative effects of different sources of variance. It is not unlikely though that the measurement instrument will be used in a manner different than it was used to estimate the G study variance components. For example, the Air Force may elect to decrease the number of tasks used on the Walk-Through Performance test from the number sampled for the G study. A D study should be conducted to assess the specific characteristics of a measurement instrument in a particular decision-making context. D studies may involve a different sample with a unique sampling of facets; oftentimes though, such a sample is not available and D study data are simulated from G study data. In either case, the D study is defined by how the organization intends to use the measurement instrument. As noted by Gillmore (1979, 1983), two critical specifications in nearly any D study are (a) the universe of generalization, and (b) the number and type of conditions to be sampled for each facet. These are explained in greater detail below.

The concept of the universe of generalization is closely related to the concept of the universe of admissible observations. The universe of admissible observations refers to the facets that one decides to include in a G study and to the range of conditions which can be sampled from each. The universe of generalization in a corresponding D study can be no larger than the universe of admissible observations; that is, it cannot contain facets missing in the universe of admissible observations, nor can it contain a broader range of conditions. The universe of generalization may be smaller, however. For example, if the universe of admissible observations included all tasks performed by jet engine mechanics, a smaller subset of tasks -- such as all cognitive tasks or all tasks using a particular tool -- may be specified for purposes of decision-making.

The more important consideration in defining a universe of generalization is deciding whether each facet is to be considered fixed or random. Random facets imply that the conditions of a facet in the D study represent a random sample from an essentially larger set of possible admissions. In practice, it is probably not necessary to actually use full random sampling procedures. However, one must at least be willing to assume that the conditions sampled in the D or G study could be replaced with other elements of some larger set of possible observations without affecting the universe score (Shavelson & Webb, 1981). For example, it is reasonable to assume that the actual administrators on any particular administration of the Walk-Through Performance Test constitute a random sample of all possible administrators, since other observers could be trained to replace them. When a random facet is specified, generalization is not limited to the set of D study conditions but instead, extends to the entire range of admissible observations.

A second possibility is that the conditions of a facet in a D study exhaust the range of conditions of interest and that generalization is intended only within that particular range. In this instance, the facet is considered fixed, and generalization is limited to the range of conditions included in the D study. Fixed effects may be treated in either of two ways. First, separate variance components for the other facets may be computed within each level of the fixed facet. For example, consider a study in which measurement occasions and rating sources (self and supervisor) were the facets of generalization. If the organization were going to use the ratings of different sources for different purposes, separate estimates of generalizability would be calculated for each source. In other instances, the organization might calculate a single summary score over all conditions of the fixed facet. For example, items on a selection test might be fixed, and each applicant receives a single total or average score. Here, the generalizability of the summary score is of interest. Since individual conditions (items) no longer matter, there can be no errors due to sampling of person-item interactions. Thus, this variance component would contribute to universe score variance and not to error variance. This would have the effect of increasing the generalizability of the test, though the improvement would hold only for the fixed set of conditions.

It should be noted that considerations of fixed and random facets occur at the D study level, not at the G study level. When computing G study estimates of variance components, all facets are treated as random (i.e., all are estimated). In subsequent D studies, the variance components are set to zero if the generalizability of average scores for that facet are of interest. Shavelson (1986) recommended inspection of the G study variance components for the fixed facet as a means of deciding how to treat the facet for D study analyses. If the variance component for the fixed facet is large, it may be inappropriate to average over its conditions and Shavelson recommends separate D study analyses for other facets within each level of the fixed facet.

Another specification in the D study is the number of conditions for each facet. These are not restricted to the number of conditions sampled in the G study. Instead, the investigator can systematically vary the number of conditions in the D study to forecast the resulting generalizability. G study variance component estimates are actually average effects for single occurrences of each facet. That is, a G study variance component represents measurement error when only a single level of the facet is used. However, measurement error decreases as scores for the object of interest are averaged over multiple levels of a facet. Analogous to the Spearman-Brown prophecy formula of classical test theory, D study estimates of variance components are computed to estimate the actual degree of error variance under varying numbers of levels of each facet. Since the variance of a mean (score over levels of the facet) is equal to the mean of the variance, D study estimates are computed by dividing the G study variance component estimates by the number of conditions specified in the D study. Logically and empirically, the error variance attributed to any one source approaches zero as the number of conditions grows infinitely large. Computing different D study estimates for differing numbers of conditions allows the researcher to predict how dependable a measure would be under a variety of measurement conditions.

Finally, it should also be noted that Cronbach et al. (1972) recognized that decision-makers use tests for different purposes; thus, they specified different error terms for each purpose. In most cases, tests are used for either relative or absolute decisions. For a relative decision, the test is used only to rank-order persons. Two common examples would be a predictor used with a top-down hiring process or a criterion which is to be correlated with a predictor. In this case, errors in persons' actual scores do not matter, as long as these errors are equivalent for each person in the sample. Problems occur only when errors exist for some persons but not for others. Thus, if a test consisted of a particularly difficult sample of items, variance due to items would not be considered error since test difficulty alone would lower all scores but preserve the rank order. However, the person-by-item variance component would be considered error since this interaction means that some persons are affected more by item difficulty than are others. In general, the error term for relative decisions includes all variance components which represent an interaction of a facet with persons.

Absolute decisions are those made about an actual score. A common example is a driving test in which the applicant must attain a certain score to receive a license. For such decisions, all effects which affect the level of the score are included in the error term. Thus, variance due to items would be included since a particularly easy or difficult sample of items would affect a person's likelihood of passing. In general, the error term for absolute decisions includes all variance components other than that which constitutes the universe score variance (typically persons). Typically, specifications of relative or absolute error terms are made at the D study level.

The discussion to this point may be summarized as follows. An investigator will conduct a single large-scale G study to estimate variance components for each effect in a model. From this single set of variance component estimates, the researcher can generate (simulate) numerous sets of D study variance component estimates and generalizability coefficients, depending upon how the measuring instrument is to be used. It is these D study results which are of the most interest to decision-makers since D study results reflect realistic or intended measurement conditions. Interpreting G and D study results is discussed further in the next section on applications of generalizability theory.

Applications of Generalizability Theory

There have been relatively few applications of generalizability theory in the field of industrial/organizational (I/O) psychology. What little has been done has been in the area of job analysis and job evaluation. Two studies (Webb & Shavelson, 1981; Webb, Shavelson, Shea, & Morello, 1981) have explored the generalizability of General Educational Development (GED) ratings by job analysts. Three other studies have applied generalizability theory to various job evaluation instruments (Doverspike & Barrett, 1984; Doverspike, Carlisi, Barrett, & Alexander, 1983; Fraser, Cronshaw, & Alexander, 1984). Two representative studies are described below.

First, Webb et al. (1981) investigated the generalizability of GED ratings of reasoning development, mathematics development, and language development by experienced job analysts. These GED ratings are subjective evaluations of the level of cognitive skill necessary to perform various jobs, and are frequently used to estimate training requirements or refer persons to job training programs. Based on in-house job descriptions, 71 analysts rated 27 jobs in terms of GED requirements. The raters were nested within one of 11 different field centers, and they rated the jobs on two different occasions. Thus, the investigators employed a three-facet design, with raters nested within offices and crossed with jobs and occasions. (Note that jobs are the object of measurement and are not considered a facet.) Separate analyses were performed for each GED rating scale.

The results showed favorable results regarding the generalizability of GED ratings. For the design described above, computed generalizability coefficients ($\varepsilon\rho^2$) of ratings from an average rater at an average center on one occasion ranged from .53 to .67 for the three scales. However, D study data showed that the generalizability coefficients ranged from .79 to .85 for the mean of four raters. Inspection of the variance components showed that differences in jobs accounted for the largest variance in ratings. The largest sources of undesirable error variance were jobs crossed with raters within centers, and the residual error term. The latter component includes random or undifferentiated error variance while the former component represents idiosyncratic perceptions of certain jobs by raters at certain centers. That is, error of this nature would result if raters at one center perceived the GED requirements of a job (such as "central-supply worker") differently than did raters at other centers.

In a second study, Fraser et al. (1984) used generalizability theory to determine the reliability of job evaluation ratings. Ratings were made by three analysts on 12 different jobs, using eight different job evaluation scales (e.g., required education, contact with others, and working conditions). Ratings were again based on in-house job descriptions. Jobs were the object of measurement, and the design was fully crossed.

The data were first analyzed with raters and scales as the facets of generalization. However, there was considerable variation due to evaluation scales so that it seemed inappropriate to assume that the different scales were simply replicates of the same facet. Instead, separate analyses were conducted for each of the eight scales. Jobs were still the object of measurement, but raters were the only facet of generalization.

For most scales, jobs accounted for the most variance in ratings. This is desirable since jobs were the object of measurement. There was little or no variance due to raters on most scales, but there was considerable variance due to the rater-by-job interaction. In other words, the results showed that jobs were differentially ranked by analysts on most scales. Still, the generalizability of job evaluation ratings was fairly high. For seven of the scales, the generalizability coefficients ranged from .70 to .92 for one random rater (G study results) and from .88 to .92 for the average of three random raters (D study results). Fraser et al. (1984) also compared their results to those of Doverspike et al. (1983), who used graduate students as raters. Interestingly, while the generalizability coefficients (and, hence, the relative size of variance components) were similar across studies, the variance components were considerably larger in the Fraser et al. field study. Presumably, this is because actual analysts used more extreme ratings to maximize discriminability among jobs.

To date, there have been no published applications of generalizability theory to performance appraisals in standard organizational settings. However, examples abound in the clinical psychology and educational measurement domains. For example, the generalizability of behavior observations or clinical assessments has been studied by Edinberg, Karoly, and Gleser (1977); Farrell, Mariotto, Conger, Curran, and Wallander (1979); Gleser, Green, and Winget (1978); Littlefield, Murrey, and Garman (1977); Mariotto and Farrell (1979); and Wieder and Weiss (1980). All applications to student evaluations of instructors are too numerous to list completely but several representative studies include: Carbno (1981); Gillmore, Kane, and Naccarato (1978); Hopkins (1983); and Kane, Gillmore, and Crooks (1976). One clinical assessment example is applicable and is described in greater detail below.

Littlefield et al. (1977) examined the generalizability of faculty ratings of third- and fourth-year dental students. Ratings were made on five general dimensions of noncognitive skills. They were collected from 31 faculty members on 12 students during one phase and from 16 faculty on 5 students during another phase. Each phase was considered a separate D study since it reflected a different set of measurement conditions (differing numbers of raters). Students were the object of measurement and the facets of generalization were raters (faculty) and rating scales. Separate D study analyses were also conducted with the scales treated as fixed or random (i.e., a mean scale score was used). The generalizability of ratings across both raters and/or scales was quite high. The generalizability coefficients were .92 and .83 for the two phases of ratings. When scales were considered fixed and the generalizability of the mean score across raters was computed, the generalizability coefficients increased to .95 and .86. In other words, about 90% of the variance in scores could be attributed to individual differences in ratees (universe score variance). Simulated D study results were also computed for a more realistic organizational condition -- ratings obtained from only one or two raters. As would be expected, generalizability coefficients were considerably lower, ranging from .53 to .61 for one rater and from .68 to .76 for two raters. Littlefield et al. concluded that at least two raters were necessary for dependable ratings.

Summary and Critique

Generalizability theory is a versatile and efficient way of estimating the effects of measurement conditions on tests and instruments. With a single design, the researcher is able to simultaneously estimate error variance due to facets such as items, test occasions, and raters. Notably, their joint effects can be assessed as well. In contrast, classical test theory enables assessment of measurement error from only one source at a time. Clearly, generalizability theory offers a more realistic portrayal of measurement error than does classical test theory.

Another primary contribution of generalizability theory is that it forces the researcher to address measurement issues that are often ignored. For example, the researcher must clearly understand the universe of admissible observations to which he or she wishes to generalize. In doing so, the researcher must explicate conditions of measurement which may affect any one individual's score. These conditions define the universe of admissible observations, place limits on the instrument's generalizability, and alert the researcher to sources of error in measurement. A second example is the D study specification of whether the instrument is to be used for relative or absolute decision-making. The degree of error differs by purpose. It is well accepted that particular tests may be dependable for grouping cases or serving as a criterion, but not for differentiating among individuals; however, there has been no acceptable way before generalizability theory to express these differences.

Finally, generalizability theory is valuable because it re-emphasizes traditional notions of reliability. Although it offers a more complex and meaningful analysis of reliability, generalizability theory is nonetheless based on the assumption that to be useful a measure must be consistent or replicable over applications. In the field of industrial/organizational psychology in general, and performance appraisal in particular, we often lose sight of the importance of reliability and consistency. Instead, issues such as rating errors, construct validity, and accuracy dominate conceptual and empirical work. Reliability is so often overlooked that few validation studies report coefficients for either predictors or criteria (e.g., Pearlman, Schmidt, & Hunter, 1980). We all know that a measure must be reliable to be valid; however, we tend to forget it. Generalizability theory refocuses our attention on reliability and offers promise for improving the reliability of measures by providing the means for pinpointing sources of error.

Analysis Plan

In that generalizability theory does have much to offer to researchers and psychometricians, a demonstration study will be presented. A broad G study will be presented along with several sets of D study results. The study examines the generalizability of performance ratings collected as part of the Air Force performance measurement project. In particular, it identifies three primary facets which could affect scores of individuals: rating forms, items or dimensions on forms, and rating sources (self, peer, or supervisor). Besides serving as a demonstration problem for the Air Force, these analyses should be of interest to other scientists as well, since the comparative value of information derived from different forms or sources has been an important research question for many years (e.g., Kraiger, 1985; Landy & Farr, 1980).

METHOD

Sample

Proficiency ratings were collected from 256 first-term U.S. Air Force enlisted personnel. All ratees were jet engine mechanics (Specialty Code 426X2). Ratees had between 6 months and 42 months of job experience and worked on one of three primary engine types.

Facets of Generalization

In the present study, there were three facets of generalization: rating forms, specific items on each form, and rating sources (self, peer, and supervisor). Items were nested within forms and both were crossed with sources. As illustrated by the Venn diagram in Figure 3, there are 12 distinct sources of variance: error, persons, sources, forms, items within forms, forms by sources, persons by sources, persons by forms, items within forms by persons, items within forms by sources, persons by forms by sources, and persons by sources by items within forms. Each facet is described in greater detail below.

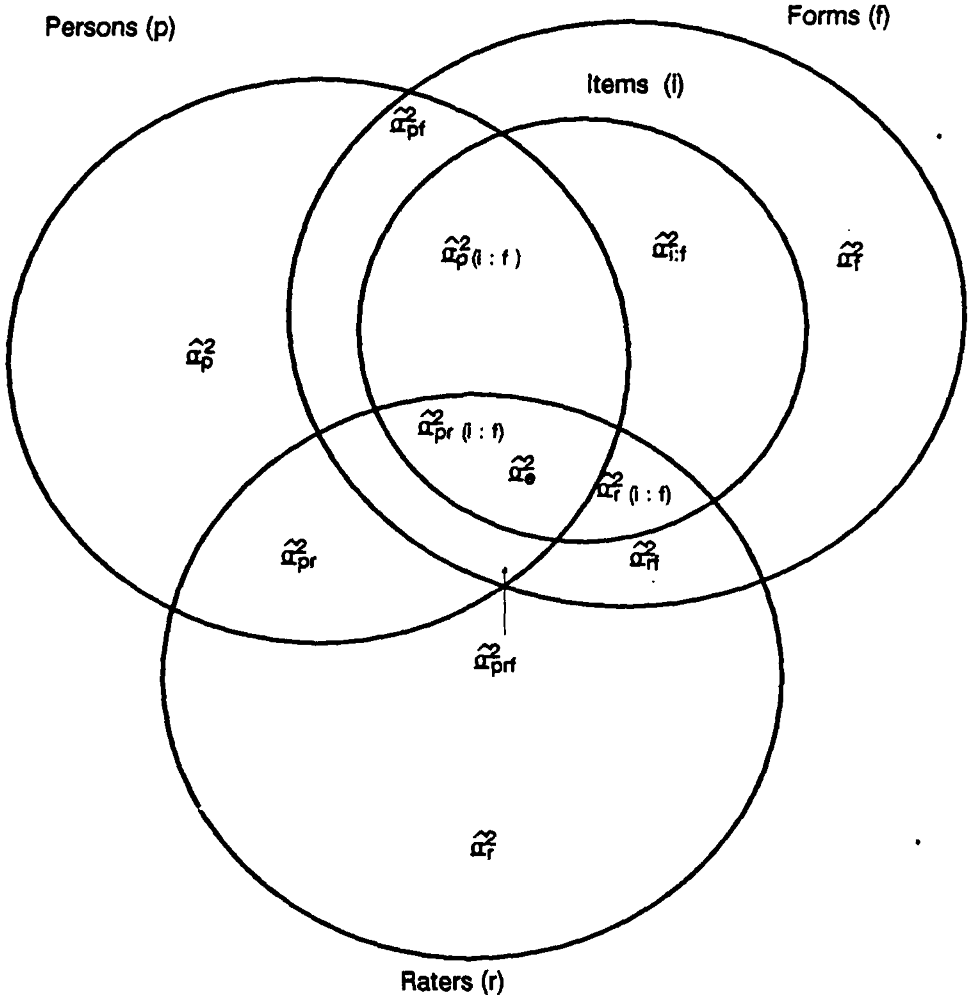

Figure 3. **Venn diagram illustrating variance components for generalizability design.**

Rating Forms. Four different forms were used to collect proficiency ratings as part of a larger Air Force research project on performance measurement. These forms can be assumed to be random samples of a larger universe of all possible forms which could be used to assess ratee performance.[1] An important research question is whether performance evaluations generalize over different rating forms.

In the present study, form content ranged from very specific to very general. The most specific was the task rating form. This form required ratings on 26 to 32 tasks representative of the job content domain. The precise number of tasks on a particular version of the form depended upon a ratee's engine assignment and

[1] An alternative assumption is that these particular forms are not random samples from the same universe but are representative of different levels within a hierarchical universe of form specificity. That is, dimensions on some forms could be nested within dimensions on more general types of forms. However, research by Vance, MacCallum, Coovert, and Hedge (1988) suggested that each form provides equivalent (and redundant) information. Moreover, a hierarchical design is less desirable in the present context because it would be more difficult to follow as a demonstration and would result in considerable loss of data because of dimensions not arranged hierarchically and because it would create an unbalanced design.

whether the principal work location was the shop or flightline. Tasks were identified from an earlier comprehensive occupational survey conducted by the Air Force. Examples of task items are: "Installs J-79 engine starter adapter pads" and "Inspects J-79 engine aircraft throttle controls for freedom of movement." Ratings were on a 5-point scale ranging from 1 (Never meets acceptable level of proficiency) to 5 (Always exceeds acceptable level of proficiency).

A slightly less specific form was the dimensional rating form. This form required ratings on six job dimensions identified through factor analyses of occupational survey data and judgments of subject-matter experts. (Again, the selection of actual dimensions on a particular version of the form depended upon a ratee's engine assignment and work location.) Examples of dimensions are: Inspect Engine, Remove/Replace Engine Components, Completion of Forms. Ratings were made on a 5-point proficiency scale similar to the one used for the task-level ratings. However, each scale point was anchored by a behavioral summary description.

More general was the global rating form. This form also used a 5-point proficiency scale anchored by behavioral descriptions; the form consisted of only two broad dimensions: technical proficiency and interpersonal proficiency. These were thought to be two important factors underlying most specific dimensions typically used in performance appraisals (Kavanagh et al., 1986).

The fourth and most general form was the Air Force-wide rating form. It required ratings on eight broad dimensions thought to be relevant to all Air Force specialties. Examples of these dimensions are: Technical Knowledge/Skill, Initiative/Effort, and Military Appearance. Ratings were again made on a 5-point proficiency scale, anchored by behavioral descriptions.

Items within Forms. A second facet of interest was individual items on each form. For each rating form, individual items or dimensions can be considered a random sample of a larger universe of all possible items which could comprise that form. However, the possible universe differs for each form. Thus, items nested within facets were considered a second facet of generalization.

There was a computational problem with this facet. In the description of the rating forms, it can be noted that the number of items or dimensions on each form ranged from 2 to 32. In standard ANOVA terms, this means that the items facet is unbalanced because there is a different number of conditions of items under each condition of forms. Though there has been little empirical work on the precise effects of unbalanced designs on variance component estimates, it is known that unbalanced designs in mixed ANOVAs often yield biased mean square estimates (Searle, 1971), which, in turn, are used to estimate variance components. In general, experts in generalizability theory recommend against the use of unbalanced designs (Brennan & Kane, 1979; Shavelson & Webb, 1981).

There are several ways of handling this problem, though none is completely desirable. One strategy would be to randomly select only two items on each form. This would be a considerable waste of data. Further, Monte Carlo work by Smith (1978) has shown that sampling errors in variance components increase substantially when each additional facet in a design has less than 7 to 10 conditions. Since sources and forms (the other two facets) already have a small number of conditions, additional estimation problems could ensue from analyzing the items facet with only two conditions.

A second strategy would be to exclude the global form with its two items, randomly select six task items and six Air Force-wide dimensions, and run the generalizability analyses with only three conditions of the form facet. This method also adds a new facet with a small number of conditions, though the error introduced should not be as severe as when a facet with only two conditions is added. Additionally, the possibility of inaccurately estimating variance due to forms increases, because this is now analyzed with only three conditions. As there is no one clearly preferable strategy, both methods were attempted and the results compared.

Rating Sources. The final facet of generalization was the source of the performance ratings. Ratings were collected from incumbents, their peers, and their supervisors. The question of interest was whether ratings generalize over sources. In some settings, these sources may be considered a random sample of a larger universe of all possible observers/raters of performance (also including subordinates, second-level supervisors, clients, etc.). However, there were no other possible rating sources of interest to the Air Force. Thus, rating sources was considered a fixed facet. Computationally, variance due to rating sources was estimated at the G study level as if sources were random. At the D study level, the rating source facet was treated in two ways. First, a mean score across sources was computed and the generalizability of this score was computed across forms and items within forms. Secondly, the generalizability of forms and items was computed for each rating source. For these analyses, there were only five sources of variance: persons, forms, items within forms, persons by forms, and persons by items within forms. Consistent with the recommendations of Shavelson (1986), it is recognized that the latter analyses would be considered the most relevant if the variance of the persons facet was relatively large.

One final note on rating sources. Mechanics were rated by one to three coworkers selected for their familiarity with the ratee and based on their availability. To again avoid the problem of an unbalanced design, a single peer rating was randomly selected for each ratee and retained for analyses.

Data Collection

Mechanics rated themselves and were rated by selected peers and supervisors. For all raters, the order of forms was global, dimensional, task, and then Air Force-wide. Ratings were made during working hours after a rigorous rater training program. All raters were informed that the ratings were being collected for research purposes.

G Study Analyses

All G and D study analyses were performed using GENOVA, a Fortran-based computer program designed for generalizability analyses (Crick & Brennan, 1983). Because GENOVA uses listwise deletion of missing data, missing data for individuals were replaced with sample means when three or fewer data points were missing; if more than three data points were missing, the entire case was deleted. As a result of the treatment of missing data, different designs employed different sample sizes. It should be noted that both analyses with and without missing data produced similar results.

For the full design, ratees (mechanics) were treated as a between-subjects factor whereas sources, forms, and items nested within forms were treated as repeated measures factors. For within-rater analyses, forms and items nested within forms were the repeated measures factors.

D Study Analyses

G study variance components were used as input for simulated D study results under different measurement conditions. All D study analyses were conducted on data from the 3-form, 6-item full G study design. These simulated D study conditions were selected to approximate current Air Force testing conditions or to represent various possible conditions of use. For example, estimated variance components and generalizability coefficients were calculated from the full design for instances in which a single rater uses a single 8-item form, or two 8-item forms, or four 8-item forms; three raters (the incumbent, a peer, and a supervisor) each use a single 8-item form; and three raters each use four 8-item forms (see Table 4). The latter condition approximates current measurement conditions on the Air Force Job Performance Measurement Project. For within-rater analyses, D study results were calculated for one 4-item, one 8-item, two 8-item, and four 2-item forms (see Tables 6, 7, and 8). Of course, D study analyses are not limited to these sets of conditions. Other decision-makers could specify their own conditions and calculate additional statistics.

Operationally, D study analyses are performed by reducing the G study variance component for each facet by the number of times the facet is measured under each particular set of measurement conditions. For example, if the estimated variance component for forms is .0016, the D study estimate would be .0008 when the number of forms is two and .0004 when the number of forms is four. Intuitively, this can be understood when it is realized that G study estimates represent average error variance for a single observation and, as in classical test theory, these estimates are reduced by a factor of the number of times they are measured.

In addition to the variance component estimates for each D study, total relative error ($\hat{\sigma}^2_\delta$) and absolute error ($\hat{\sigma}^2_\Delta$) were calculated, along with a corresponding generalizability coefficient. The generalizability coefficient represents the proportion of universe score variance to total variance. For relative decisions (e.g., for validation purposes), total variance is equal to universe score variance plus total relative error. This coefficient represents the generalizability of ratings over conditions of measurement when all facets are random. If a facet were to be considered fixed, a new generalizability coefficient could be calculated after first adding to the universe score variance all D study variance components involving the facet.

RESULTS

Generalizability results will be presented as follows. First will be G study analyses for the full design (raters by items within forms) and within-rater designs (items nested within forms). For the full design, separate results are presented for analyses with six items within three forms, and two items within four forms. Only three-form analyses are presented for the within-rater analyses. Next, simulated D study results are presented for the full design and within-rater designs.

G Study: Full Design

G study estimates of variance components with 90% confidence intervals are presented in Tables 1 and 2. The confidence intervals indicate the precision in estimation of the population values of variance components, given the sample size and design complexity. The confidence intervals are based on the ratio of the estimated variance component to its standard error and were calculated from procedures detailed by Satterthwaite (1941, 1946). Satterthwaite's method corrects the upper limit of the interval which is frequently too low when calculated as the product of the normal deviate and the standard error. Also included in the tables are degrees of freedom and mean squares associated with each effect. It should be noted that the variance components are very similar across designs. One exception is variance due to forms. Not surprisingly, by increasing the number of forms in the G study design from three to four, variance due to forms is decreased (from .033 to .001). The confidence intervals in Table 1 are generally narrower than those of Table 2, reflecting smaller standard errors and greater degrees of freedom in the three-form, six-item analysis.

Inspection of the variance components in either table reveals several interesting findings. First, the most variance is attributed to the residual term, $\hat{\sigma}^2_{pr(i:f)}$ (.290 and .293 in the six items within three forms and the two items within four forms designs, respectively). This represents undifferentiated error. The effect for persons, variance due to individual differences, was fairly large in both designs ($\hat{\sigma}^2_p$ = .100, .151). This variance is desirable and represents universe score variance. Among the other error sources of variance in the design, the larger effects involved the interaction of rater sources and ratees. $\hat{\sigma}^2_{pr}$ (.278, .186) was large, indicating at least two sources differentially ranked ratees. $\hat{\sigma}^2_{prf}$ (.053, .016) was also nontrivial, indicating that sources differentially ranked ratees, but did so differently on different forms.

It should be recalled that rating sources can be considered a fixed facet for subsequent D study analyses. After Shavelson (1986), a fixed facet can be treated by averaging scores across levels of the facet or analyzing other facets within levels of the facet. The generalizability of scores over levels of rater sources was investigated and these results are presented in Table 3.

Table 1. **Estimated Variance Components for G Study with Six Items and Three Forms**

Effect	DF	MS	$\hat{\sigma}^2$	90% confidence intervals
Persons (p)	222	11.550	.100	.072 < $\hat{\sigma}^2$ < .148
Raters (r)	2	89.420	.020	.011 < $\hat{\sigma}^2$ < .060
Forms (f)	2	152.005	.033	.017 < $\hat{\sigma}^2$ < .098
Items w/in Forms (i:f)	15	15.846	.022	.013 < $\hat{\sigma}^2$ < .046
pr	42	5.602	.278	.246 < $\hat{\sigma}^2$ < .316
pf	442	1.172	.022	.016 < $\hat{\sigma}^2$ < .033
rf	4	2.900	.001	.001 < $\hat{\sigma}^2$ < .003
prf	84	.606	.053	.045 < $\hat{\sigma}^2$ < .062
p(i:f)	3,315	.460	.057	.051 < $\hat{\sigma}^2$ < .064
r(i:f)	30	.892	.003	.002 < $\hat{\sigma}^2$ < .006
pr(i:f)	6,630	.290	.290	.282 < $\hat{\sigma}^2$ < .298

Table 2. **Estimated Variance Components for G Study with Two Items and Four Forms**

Effect	DF	MS	$\hat{\sigma}^2$	90% confidence intervals
Persons (p)	206	5.591	.151	.119 < $\hat{\sigma}^2$ < .199
Raters (r)	2	27.841	.015	.008 < $\hat{\sigma}^2$ < .044
Forms (f)	3	12.147	.001	.001 < $\hat{\sigma}^2$ < .003
Items w/in Forms (i:f)	4	10.639	.015	.008 < $\hat{\sigma}^2$ < .044
pr	412	1.812	.186	.163 < $\hat{\sigma}^2$ < .214
pf	618	.477	.000	.000 < $\hat{\sigma}^2$ < .000
rf	6	1.434	.001	.000 < $\hat{\sigma}^2$ < .002
prf	1,236	.326	.016	.009 < $\hat{\sigma}^2$ < .048
p(i:f)	824	.464	.057	.046 < $\hat{\sigma}^2$ < .074
r(i:f)	8	1.086	.004	.002 < $\hat{\sigma}^2$ < .011
pr(i:f)	1,648	.293	.293	.276 < $\hat{\sigma}^2$ < .311

G Study: Within-Rater Design

The relatively large effect for the persons-by-source interaction indicates that ratees are differentially ranked by sources. This finding suggests that it may be inappropriate to average scores over sources and that separate analyses for other facets should be conducted within rater sources. Estimated variance components of an items within forms analysis for each rating source are presented in Table 3. Results presented are based on six items within three forms; similar results (not presented) were obtained from analyses on two items within four forms. The table also contains degrees of freedom, mean squares, and 90% confidence intervals associated with each effect.

Table 3. **Study Variance Components Within Rater Sources**

Effect	DF	MS	$\hat{\sigma}^2$	90% confidence intervals
Self:				
Persons (p)	217	4.127	.193	$.161 < \hat{\sigma}^2 < .235$
Forms (f)	2	51.736	.034	$.018 < \hat{\sigma}^2 < .102$
Items w/in forms (i:f)	15	5.865	.025	$.015 < \hat{\sigma}^2 < .052$
pf	434	.666	.053	$.042 < \hat{\sigma}^2 < .068$
pi:f	3,255	.351	.351	$.337 < \hat{\sigma}^2 < .368$
Supervisor:				
Persons (p)	217	7.677	.375	$.317 < \hat{\sigma}^2 < .453$
Forms (f)	2	41.104	.026	$.014 < \hat{\sigma}^2 < .077$
Items w/in forms (i:f)	15	6.064	.026	$.016 < \hat{\sigma}^2 < .054$
pf	434	.926	.097	$.082 < \hat{\sigma}^2 < .117$
pi:f	3,255	.346	.346	$.333 < \hat{\sigma}^2 < .360$
Peer:				
Persons (p)	217	5.572	.265	$.222 < \hat{\sigma}^2 < .321$
Forms (f)	2	67.000	.047	$.024 < \hat{\sigma}^2 < .137$
Items w/in forms (i:f)	15	5.481	.024	$.014 < \hat{\sigma}^2 < .049$
pf	434	.809	.077	$.064 < \hat{\sigma}^2 < .094$
pi:f	3,255	.350	.350	$.336 < \hat{\sigma}^2 < .364$

Across rating sources, the largest variance is generally attributed to undifferentiated error, $\hat{\sigma}^2_{p(i:f)}$. The size of this estimated variance component was consistent across sources. Other sources of error variance were small relative to effects for undifferentiated error and individual differences. The forms and persons-by-forms effects are both somewhat larger for supervisory and peer ratings than for self ratings.

The relative size of the variance component for persons is much higher in these analyses than in the full design. This is expected since the universe of admissible observations is smaller. In other words, scores of persons are more generalizable within a smaller domain. Examining the relative size of variance components across sources, it can be seen that universe variance is smallest for self ratings ($\hat{\sigma}^2_p = .193$) and largest for supervisory ratings ($\hat{\sigma}^2_p = .375$). Overall, the relative proportion of design variance attributed to $\hat{\sigma}^2_p$ is larger in the within-rater design than in the full design.

D Studies: Full Design

Results of simulated D study analyses of the full design are presented in Tables 4 and 5. For the analyses presented in Table 4, rater source was assumed to be a random facet and five different specifications of measurement conditions were made: one rater, one 8-item form; one rater, two 8-item forms; one rater, four 8-item forms; three raters, one 8-item form; three raters, four 8-item forms. For analyses presented in Table 5, the rater source facet was considered fixed (with three conditions) and four measurement specifications were made: one 8-item form, two 4-item forms, four 8-item forms, and one 12-item form. Assuming that sources are fixed implies that the specific rater types sampled at the G study level exhaust the universe of possible rater sources and that individuals' scores are standardized (summed or averaged) over rating sources.

Table 4. **Simulated D Study Results of Full Design, Illustrating Changes in Raters or Forms**[a]

$\hat{\sigma}^2$ for pr (i:f) design		$\hat{\sigma}^2$ for pR (I:F) design				
	n_r	1	1	1	3	3
	n_f	1	2	4	1	4
	n_i	8	8	8	8	8
$\hat{\sigma}^2_p = .100$	$\hat{\sigma}^2_p =$	.100	.100	.100	.100	.100
$\hat{\sigma}^2_r = .020$	$\hat{\sigma}^2_R =$	.020	.020	.020	.007	.007
$\hat{\sigma}^2_f = .033$	$\hat{\sigma}^2_F =$	.033	.017	.008	.033	.008
$\hat{\sigma}^2_{i:f} = .022$	$\hat{\sigma}^2_{I:F} =$	.003	.001	.001	.003	.001
$\hat{\sigma}^2_{pr} = .278$	$\hat{\sigma}^2_{pR} =$	.278	.278	.278	.093	.093
$\hat{\sigma}^2_{pf} = .022$	$\hat{\sigma}^2_{pF} =$	.022	.011	.006	.022	.006
$\hat{\sigma}^2_{rf} = .001$	$\hat{\sigma}^2_{RF} =$	.001	.001	.000	.000	.000
$\hat{\sigma}^2_{prf} = .053$	$\hat{\sigma}^2_{pRF} =$	.053	.026	.013	.018	.004
$\hat{\sigma}^2_{p(i:f)} = .057$	$\hat{\sigma}^2_{p(I:F)} =$	.007	.004	.002	.007	.002
$\hat{\sigma}^2_{r(i:f)} = .003$	$\hat{\sigma}^2_{R(I:F)} =$	.000	.000	.000	.000	.000
$\hat{\sigma}^2_{pr(i:f)} = .290$	$\hat{\sigma}^2_{pR(I:F)} =$	.036	.018	.009	.012	.003
	$\hat{\sigma}^2_\tau =$	.100	.100	.100	.100	.100
	$\hat{\sigma}^2_\delta =$	.395	.337	.307	.151	.107
	$\hat{\sigma}^2_\Delta =$	.454	.376	.337	.195	.123
	$\hat{\sigma}^2_X =$	.129	.436	.407	.251	.207
	$E\rho^2 =$	.201	.229	.245	.397	.482
	$\Phi =$	.180	.210	.228	.338	.447

[a]Based on G study results for 3-form, 6-item analysis.

The notation used to present the D study results is drawn from Brennan (1983; Brennan & Kane, 1979) and requires some explanation. Whereas lowercase letters are used to present G study variance components (e.g, $\hat{\sigma}^2_f$ and $\hat{\sigma}^2_{pr}$), uppercase letters represent the corresponding D study estimates (e.g., $\hat{\sigma}^2_F$ and $\hat{\sigma}^2_{pR}$). Thus, a capital letter for a facet (recall that persons are not a facet) signifies that the variance component has been averaged over the levels of the effect indicated by the D study measurement conditions. For example, with three raters and four forms, the estimated variance component $\hat{\sigma}^2_{rf}$ is averaged over 12 cells and $\hat{\sigma}^2_{RF} = \hat{\sigma}^2_{rf}/(12)$.

The lower portions of Tables 4 and 5 present estimates of universe score variance ($\hat{\sigma}^2_\tau$), relative error variance ($\hat{\sigma}^2_\delta$), absolute error variance ($\hat{\sigma}^2_\Delta$), total observed score variance ($\hat{\sigma}^2_X$), and two generalizability coefficients ($E\rho^2$ and Φ). Exact computational formulas and theoretical explanations for these values are given in Brennan (1983). Briefly, universe score variance equals $\hat{\sigma}^2_p$. In a random model, the relative error variance is equal to the sum of all effects which contain p and at least one other index (e.g., $\hat{\sigma}^2_{pf}$) and the absolute error variance is equal to the sum of all effects in the design except $\hat{\sigma}^2_p$. Total observed score variance is equal to universe score variance plus relative error variance. The generalizability coefficient $E\rho^2$ is equal to the ratio of $\hat{\sigma}^2_\tau$ to the sum of universe score variance plus relative error variance, and Φ is equal to the ratio of $\hat{\sigma}^2_\tau$ to the sum of universe score variance plus absolute error variance.

Looking first at Table 4, it can been seen that in comparison to the G study results, measurement error is reduced considerably by replications of raters, forms, or items. For example, undifferentiated error variance ($\hat{\sigma}^2_{pR(I:F)}$) drops from .290 to .036 by simply averaging over eight items on a single form with a single rater. This set of conditions also decreases variance in all other effects which contain the item facet ($\hat{\sigma}^2_{I:F}$ and $\hat{\sigma}^2_{pR:F}$. Measurement specifications which contain more than one rater or form condition result in reductions in error variance for effects containing either the rater or form facet. For example, variance due to forms decreases from .033 when one form is used to .008 when four forms are used. When scores are averaged over three sources, four forms, and eight items, there is virtually no error variance due to items within forms, the source-by-form interaction, the person-by-items-within-forms interaction, or the sources by items within forms interaction.

Table 5. **Simulated D Study Results of Full Design, Illustrating Sources Fixed, Illustrating Changes in Items**[a]

$\hat{\sigma}^2$ for pr (I:f) design		$\hat{\sigma}^2$ for pR (I:F) design			
	n_f	1	2	4	1
	n_i	8	4	8	12
$\hat{\sigma}^2_p = .100$		$\hat{\sigma}^2_p = .100$	.100	.100	.100
$\hat{\sigma}^2_r = .020$		$\hat{\sigma}^2_R =$ [b]			
$\hat{\sigma}^2_f = .033$		$\hat{\sigma}^2_F = .033$	.017	.008	.033
$\hat{\sigma}^2_{i:f} = .022$		$\hat{\sigma}^2_{I:F} = .003$	.003	.001	.002
$\hat{\sigma}^2_{pr} = .278$		$\hat{\sigma}^2_{pR} =$ [b]			
$\hat{\sigma}^2_{pf} = .022$		$\hat{\sigma}^2_{pF} = .022$	.011	.006	.022
$\hat{\sigma}^2_{rf} = .001$		$\hat{\sigma}^2_{RF} =$ [b]			
$\hat{\sigma}^2_{prf} = .053$		$\hat{\sigma}^2_{pRF} =$ [b]			
$\hat{\sigma}^2_{p(I:f)} = .057$		$\hat{\sigma}^2_{p(I:F)} = .007$	.007	.002	.005
$\hat{\sigma}^2_{r(I:f)} = .003$		$\hat{\sigma}^2_{R(I:F)} =$ [b]			
$\hat{\sigma}^2_{pr(I:f)} = .290$		$\hat{\sigma}^2_{pR(I:F)} =$ [b]			
		$\hat{\sigma}^2_\tau = .100$	.100	.100	.100
		$\hat{\sigma}^2_\delta = .029$	.018	.007	.027
		$\hat{\sigma}^2_\Delta = .065$	.038	.016	.062
		$\hat{\sigma}^2_X = .129$	.118	.107	.126
		$E\rho^2 = .774$	.846	.932	.789
		$\Phi = .604$	.726	.859	.616

[a] Based on G study results for 3-form, 6-item analysis.

[b] No variance since scores are averaged over sources.

The lower half of Table 4 presents universe variances, two error variances, total observed score variances, and generalizability coefficients for each set of measurement conditions. The universe score variance ($\hat{\sigma}^2_\tau$) remains the same regardless of conditions. It can be seen that when rater sources are considered random, the generalizability of ratings is fairly low regardless of the measurement conditions. $E\rho^2$ (the generalizability for relative decisions) ranges from .201 when one rater uses one 8-item form to .482 with scores averaged over three raters using four 8-item forms. Φ, the generalizability coefficient for absolute decisions, is somewhat lower. Since the estimated variance component for the persons-by-rater interaction is large, it appears that all three rating sources are necessary conditions to maximize generalizability.

A more positive conclusion emerges from inspection of results in Table 5, based on the assumption of fixed rater sources. In these analyses, ratings are averaged over the three sources, and the generalizability of these means over the facets of interest are investigated. Because scores are averaged over sources, there can be no variance due to different sources and variance components for effects which contain sources are set to zero for the D study analyses.

When rater sources are assumed to be fixed, the mean rating over sources is very generalizable under a variety of conditions. $E\rho^2$, the generalizability coefficient for relative decisions, ranges from .774 with one 8-item form to .932 with four 8-item forms. Corresponding values for ϕ are somewhat lower. Generalizability coefficients above .70 are generally considered adequate for decision-making purposes. It is interesting to note that while adding more items is considerably easier than adding more rating forms, it has less of an effect on the generalizability of ratings. For example, adding four additional items to the single 8-item form will only increase the generalizability coefficient from .774 to .789, but changing from a single 8-item form to two 4-item forms instead will increase the coefficient to .846 (see Table 5).

D Studies: Within-Raters

Results of similar analyses for self, supervisory, and peer ratings are presented in Tables 6, 7, and 8. Measurement conditions were specified as one 4-item form, one 8-item form, two 8-item forms, and four 2-item forms. Results are presented for only the 3-form, 6-item G study analysis, though results were comparable for both designs. Comparing the results for the three rating sources, it appears that ratings from supervisors are more dependable over forms and items than are ratings from incumbents or peers. Ratings by peers are also more generalizable than ratings by incumbents (self ratings). However, regardless of source, ratings appear to be fairly generalizable within source when at least two 8-item forms are used. The generalizability coefficients under these conditions for self, supervisory, and peer ratings are .800, .843, and .815, respectively. Coefficients for ratings over four 2-item forms are not appreciably different from those for ratings over two 8-item forms.

Inspection of the D study variance components for all sources reveals that the persons-by-forms effect is the largest source of error variance. While this effect is reduced by increasing the number of forms, it remains relatively large in comparison to other sources. This effect represents variance due to ratees being differentially ranked on different forms. Attention to this problem, perhaps during rater training, could result in more generalizable ratings.

Table 6. Simulated D Study Results of Self Ratings[a]

		$\hat{\sigma}^2$ for p (I:F) design			
$\hat{\sigma}^2$ for p (i:f) design	n_f	1	1	2	4
	n_i	4	8	8	2
$\hat{\sigma}^2_p = .192$	$\hat{\sigma}^2_p =$	.192	.192	.192	.192
$\hat{\sigma}^2_f = .035$	$\hat{\sigma}^2_F =$	.035	.035	.017	.009
$\hat{\sigma}^2_{i:f} = .025$	$\hat{\sigma}^2_{I:F} =$	.006	.003	.002	.003
$\hat{\sigma}^2_{pf} = .053$	$\hat{\sigma}^2_{pF} =$	.053	.053	.026	.014
$\hat{\sigma}^2_{p(i:f)} = .351$	$\hat{\sigma}^2_{p(I:F)} =$	.088	.044	.022	.044
	$\hat{\sigma}^2_\tau =$	.192	.192	.192	.192
	$\hat{\sigma}^2_\delta =$	.140	.096	.048	.057
	$\hat{\sigma}^2_\Delta =$	.181	.134	.067	.069
	$\hat{\sigma}^2_X =$	.332	.289	.240	.248
	$E\rho^2 =$	.578	.666	.800	.774
	$\phi =$	.515	.589	.741	.738

[a]Based on G study results for 3-form, 6-item analysis.

Table 7. **Simulated D Study Results of Supervisory Ratings**[a]

$\hat{\sigma}^2$ for p (i:f) design		$\hat{\sigma}^2$ for p (I:F) design			
	n_f	1	1	2	4
	n_i	4	8	8	2
$\hat{\sigma}^2_p = .375$	$\hat{\sigma}^2_p =$	.375	.375	.375	.375
$\hat{\sigma}^2_f = .026$	$\hat{\sigma}^2_F =$	.026	.026	.013	.007
$\hat{\sigma}^2_{i:f} = .026$	$\hat{\sigma}^2_{I:F} =$	.007	.003	.002	.003
$\hat{\sigma}^2_{pf} = .097$	$\hat{\sigma}^2_{pF} =$	.097	.097	.048	.024
$\hat{\sigma}^2_{p(i:f)} = .346$	$\hat{\sigma}^2_{p(I:F)} =$	.086	.043	.022	.043
	$\hat{\sigma}^2_\tau =$	.375	.375	.375	.375
	$\hat{\sigma}^2_\delta =$	.183	.140	.070	.068
	$\hat{\sigma}^2_\Delta =$	.216	.170	.085	.077
	$\hat{\sigma}^2_X =$	.558	.515	.445	.443
	$E\rho^2 =$	.671	.728	.843	.847
	$\Phi =$	.634	.689	.816	.829

[a] Based on G study results for 3-form, 6-item analysis.

Table 8. **Simulated D Study Results of Peer Ratings**[a]

$\hat{\sigma}^2$ for p (i:f) design		$\hat{\sigma}^2$ for p (I:F) design			
	n_f	1	1	2	4
	n_i	4	8	8	2
$\hat{\sigma}^2_p = .265$	$\hat{\sigma}^2_p =$	.265	.265	.265	.265
$\hat{\sigma}^2_f = .047$	$\hat{\sigma}^2_F =$	.047	.047	.023	.012
$\hat{\sigma}^2_{i:f} = .024$	$\hat{\sigma}^2_{I:F} =$	.006	.003	.001	.003
$\hat{\sigma}^2_{pf} = .077$	$\hat{\sigma}^2_{pF} =$	.077	.077	.038	.019
$\hat{\sigma}^2_{p(i:f)} = .350$	$\hat{\sigma}^2_{p(I:F)} =$	.087	.044	.022	.044
	$\hat{\sigma}^2_\tau =$	.265	.265	.265	.265
	$\hat{\sigma}^2_\delta =$	.164	.120	.060	.063
	$\hat{\sigma}^2_\Delta =$	.217	.170	.085	.077
	$\hat{\sigma}^2_X =$	.429	.385	.325	.306
	$E\rho^2 =$	.617	.688	.815	.808
	$\Phi =$	.550	.609	.757	.774

[a] Based on G study results for 3-form, 6-item analysis.

DISCUSSION

The format of this discussion section will be as follows. First, the results will be interpreted within the context of both generalizability theory and performance appraisal. This will be followed by an assessment of the relevancy of generalizability theory to the project and some final conclusions and recommendations.

Interpretation of Results

The results indicate that performance ratings collected as part of the performance measurement project are somewhat generalizable across several different universes of generalization. As would be expected, ratings are more generalizable within smaller universes of admissible observations (e.g.. rating sources). For the full design, ratings will have an acceptable level of generalizability only when the three rating sources are assumed fixed and at least two forms are used. Under these conditions, the generalizability coefficient for ratings exceeds .80, which is the minimum acceptable level proposed by Cardinet, Tourneur, and Allal (1976). There are several ways of interpreting this generalizability coefficient. Literally, it is the ratio of universe score variance to observed score variance and is analogous to the reliability coefficient in classical test theory. Alternatively, it may be understood as an intraclass correlation representing the average correlation of observed deviation scores (from their overall mean) and universe deviation scores (from their overall mean). Perhaps the most straightforward interpretation is that the generalizability coefficient represents the proportion of observed score variance which can be attributed to the object being measured. Thus, when ratings are averaged over three rater sources each using four 8-item forms, nearly 90% of the observed score variance is due to individual differences, and only 10% is due to measurement error. Since the forms and items within forms facets were considered random for this analysis, other rating forms or items could be sampled from the same universe of admissible observations with no change in the generalizability coefficient.

For D study analyses within rater level, somewhat smaller sets of measurement conditions are needed to achieve acceptable generalizability coefficients. Only two 8-item forms are needed for supervisory and peer ratings to produce generalizability coefficients greater than .80. As the number of items or forms is increased further, generalizability coefficients for supervisory and peer ratings may exceed .90. It should be noted that while these coefficients are higher than for the full design, the universe of generalizability is smaller, as it is limited to only a single rater source.

For all analyses, there are several notable results involving individual variance components. The largest variance component in each G study was the residual term. This term represents the confounding of the most complex interaction term and undifferentiated error. Presumably, the size of this term could be decreased by identifying other potential sources of measurement error (e.g., measurement occasions) and including them in the design. Although this strategy does not reduce total error variance (it only reapportions the error variance), it does serve to explicitly identify each source. Once the degree of error due to different facets is known, it can be controlled in future testing administrations through more careful measurement procedures or by averaging scores across conditions of the facet.

It should also be noted that there is very little error variance due to sampling of forms. This finding is very similar to results of covariance structural modeling analyses by Vance et al. (1988), who reported that the traits measured by different forms are nearly perfectly correlated when measurement error is controlled for. The present data are more persuasive though, since Vance et al. investigated only relationships of means over all form items. The present analyses included items as a separate facet, yet still found no effect for forms alone. These results also substantiate the conclusions of Jacobs, Kafry, and Zedeck (1980), who reported that different rating forms had little effect on measurement quality.

In general, the present results also compare favorably to those of other studies of generalizability theory and performance appraisal. In a generalizability study of faculty ratings of dental students, Littlefield et al.

(1977) reported generalizability coefficients in the .70's for two raters and five items. Two separate studies of student evaluations of teachers reported by Kane et al. (1976) yielded generalizability coefficients between .70 and .82 for 10 items and 10 raters. In another study of course evaluations by Gillmore et al. (1978), the generalizability of scores over 5 raters and 10 items was only .59. Thus, the present results are at least as high or higher than many generalizability coefficients typically reported in the literature.

Logical Requirements for Performance Ratings

Dickinson (1986) demonstrated how random effects in an analysis of variance design can be interpreted in terms of logical requirements for performance appraisal. This perspective can be applied to the present results as well. One primary extension of generalizability theory beyond Dickinson's analyses is that variance components can be assessed at the D study level as well as for G studies (Dickinson's level of analysis).

After Lawler (1967), Kavanagh, MacKinney, and Wolins (1971), and others, Dickinson (1986) argued that performance ratings should possess validity as do other measures of individual differences. Specifically, ratings should be shown to have high convergent validity (among methods or sources), moderately high discriminant validity (across rating dimensions), and low method bias (i.e., method of rating affects ratee ordering). Dickinson then demonstrated how indices of these logical requirements can be calculated through analysis of variance applied to multiple measures of ratees.

Convergent validity is indicated by the variance component for those effects which interact with persons (e.g., $\hat{\alpha}_{pf}$ or $\hat{\alpha}_{pr}$). These variance components indicate the degree to which ratees are ordered invariantly over different forms or rating sources. The $\hat{\alpha}_{pf}$ is relatively small, indicating convergent validity over rating forms. In contrast, $\hat{\alpha}_{pr}$ is relatively high, indicating a lack of convergence between rating sources. Generalizability coefficients within sources are satisfactory, suggesting that ratings are dependable (or reliable) within sources, but not across sources.

Discriminant validity is indicated by the variance component for the interaction of persons and items within forms ($\hat{\alpha}_{pi:f}$). Effective raters recognize individual strengths and weaknesses in ratees and discriminate among these attributes when making ratings. In the present analyses, $\hat{\alpha}_{pi:f}$ was moderately high for both the 6-item, 3-form and the 2-item, 4-form G study analyses, suggesting reasonable evidence of discriminant validity. A similar analysis could not be made within rater sources as the variance component $\hat{\alpha}_{p(i:f)}$ contains both the person by items within forms effect and undifferentiated error.

Assessment of Generalizability Theory

The present application illustrated several strengths and weaknesses of generalizability theory. On the positive side, it showed how a multifaceted approach to measurement error can aid decision-makers in refining an instrument. Specifically, it was found in D study analyses of the full design that variance due to the ratee-rater and ratee-rater-form interactions were fairly large, even when averaging across three rater sources. Because this is a multivariate approach, it is known that this error exists independent of variance due to forms or items on forms. If the Air Force considers these to be undesirable types of error variance, these errors can be treated either statistically by increasing the number of rater sources (though this might not be practical) or methodologically by an intervention in the instrument development or administration stage. For example, the rater training process could be altered.

Another advantage is that generalizability theory yields indices of rater dependability in situations in which classical test theory approaches may be unable to do so. For example, the primary index of rater quality in classical theory is an assessment of interrater agreement, but this requires at least two raters. In situations in which only a single rater exists, generalizability theory can still generate other measures of rater dependability in terms of forms, items, occasions, etc.

A very important practical application is that generalizability theory allows decision-makers to assess the generalizability of an instrument under conditions other than the ones in current use. In the present study, the only analysis of the full design which could be run involved four items on two forms. In all likelihood, the Air Force would never actually collect ratings under these conditions. Yet this analysis provided trustworthy estimates of variance components under a variety of conditions involving larger numbers of forms or items. This information could be very useful for decisions about altering the current assessment process. Ideally, information from a generalizability study could be combined with utility data to make rational decisions about how many raters, forms, or items should be used. For example, simulated D study results for the full design revealed that all three rating sources are necessary to ensure dependable ratings. However, there is not a great gain in generalizability from increasing the number of items. Utility analyses could be used to select the most cost-efficient form, and rater time demands could be lessened by using only a single form.

A final advantage comes from the process of conducting a generalizability study. As mentioned above, generalizability theory forces researchers and decision-makers to explicitly address measurement issues which are too often ignored. For example, what are all the conditions of measurement which could affect observations of individuals? How can these be measured and controlled? Are the measurement instruments to be used for relative or absolute decision-making? Are particular measurement conditions to be considered random samples of a larger set of possible conditions or do they exhaust the set? Will the same set of conditions always be used, or might a smaller or larger set be used in the future? Perhaps the most valuable contribution is that generalizability theory refocuses our attention on the reliability of our measures and reminds us that without this basic property, our measures cannot be valid for other purposes.

One potential problem is the accuracy of estimated variance components. Indeed, Shavelson and Webb (1981) called sampling errors in variance components the "Achilles' heel" of generalizability theory. In the present study, some confidence intervals were large enough to include zero due to the size of their standard errors. Proposed solutions to the sampling error problem include large samples, large numbers of conditions for each facet, and the use of multiple designs (Smith, 1978). For the Air Force, there should not be a problem in obtaining large samples for many occupational specialties, but increasing the number of conditions would probably be impossible since the instruments are already designed. Moreover, it would be infeasible to give raters nine or ten 20-item forms just to ensure more accurate estimates of variance components. The third solution is more workable though, since the Air Force will be collecting identical rating data from additional occupational specialties. The same analyses should be repeated on each specialty and the results compared. Then, variance components could be averaged across specialties for more accurate estimates of population parameters.

More importantly, the question must be raised again as to whether the objectives and benefits of generalizability theory are consistent with the current directions of the Air Force performance measurement project. According to Kavanagh et al. (1986), the Air Force will assess the quality of rating data primarily in terms of construct validity and accuracy. This should include results of generalizability theory which can be interpreted as evidence of convergent validity or discriminant validity. However, considerable other evidence (e.g., content validity, correlations with nonrating data) must be accumulated before fully informed decisions can be made about the construct validity of these measures. Further, without known target scores included in a design, generalizability analyses would have little relevancy to the accuracy of the rating measures. The application of generalizability theory to questions of accuracy when such scores are available is an interesting problem and is discussed below.

Generalizability theory is first and foremost a useful method for assessing the dependability of scores over conditions of measurement. It is probably most useful if used early in the instrument development stage when decision-makers still have some latitude in determining future measurement conditions. However, it can also be useful after the instrument is in operation, for modifying or optimizing existing procedures. For example, the present study strongly demonstrates the need to retain all three rating sources in order to maximize the generalizability of the ratings. As discussed in the introduction, if an instrument cannot be shown to be test reliable or dependable over measurement conditions, questions of validity or accuracy are moot.

RECOMMENDATIONS

1. It is important that the performance measurement project continue to collect ratings from all three rating sources. Levels of generalizability are unacceptably low across sources unless data from all three rating sources are collected and averaged.

2. Generalizability analyses should be applied to other occupational specialties, permitting the comparison of estimated variance components and generalizability coefficients across specialties. If specialties were a facet, the variance due to specialties could be specified. If the variance component for specialties were small, variance components for other effects could be combined over specialties for more accurate estimates of population parameters. Alternatively, Monte Carlo studies could be performed to identify the total number of designs (or subjects across designs) needed to compute accurate population estimates.

3. Generalizability analyses can also be applied to the Walk-Through Performance Test component of the Job Performance Measurement Project. This would seem to be a fruitful endeavor since there are currently tremendous manpower costs due to the length of the tests. Generalizability theory could be used to assess changes in the dependability of scores with reductions in the number of administrators, tasks, dimensions, or tests themselves (i.e, elimination of the interview or hands-on component).

4. The application of generalizability theory to questions of construct validity could be explored further by investigating the consistency of scores over different types of performance measures (i.e., ratings, hands-on, interviews). This application of generalizability theory has been recommended by Morse and Morse (1976), who also recommended using binary pass/fail scores on each criterion. Evidence of construct validity would come from the percentage of incumbents who "pass" the walk-through and also receive "passing" proficiency ratings. Alternatively, data from the Walk-Through and performance ratings could be used to investigate questions of rater accuracy. If the Walk-Through is truly a benchmark, then accurate raters are those whose ratings most closely approximate these ratings. Dickinson (1986) has also shown how analysis of variance designs can be used to draw conclusions about the accuracy of performance ratings. Similar conclusions can be reached through generalizability theory. For example, if supervisory ratings and Walk-Through scores are compared, a small variance component for sources could be interpreted as what Cronbach (1955) termed "elevation accuracy," or the extent of agreement between the ratings and the target scores.

5. Information about the generalizability of measures should be combined with other information in making decisions about the usefulness of various criterion measures. For example, the cost of each measure can be expressed as a function of the time and expense necessary to develop the measure and/or collect data using the measure. The benefit of any set of measures can be expressed as a function of their generalizability levels and any possible attenuating effects on the relationship of these measures to the Armed Services Vocational Aptitude Battery (ASVAB). These data can be used to derive informed answers to questions about criterion measures, such as: Which combination of criterion measures can be collected the most inexpensively without adversely affecting our ability to validate the ASVAB?

REFERENCES

Brennan, R.L. (1983). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Brennan, R.L., & Kane, M.T. (1979). Generalizability theory: A review. In L.J. Fryans, Jr. (Ed.), *Generalizability theory: Inferences and practical applications.* San Francisco: Jossey-Bass.

Carbno, W.C. (1981). Student evaluation of teacher effectiveness: A case study. *Educational and Psychological Measurement,* 41, 937-951.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement,* 13, 119-135.

Crick, J.E., & Brennan, R.L. (1983). *GENOVA: A generalized analysis of variance program (FORTRAN IV computer program and manual).* Dorchester, MA: Computer Facilities, University of Massachusetts.

Cronbach, L.J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin,* 52, 177-193.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements.* New York: Wiley.

Cronbach, L.J., Rajaratnam, N., & Gleser, B. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology,* 16, 137-163.

Dickinson, T.L. (1986, July). *Performance ratings: Designs for evaluating their validity and accuracy* (AFHRL-TP-86-15, AD-A170 400). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Doverspike, D., & Barrett, G.V. (1984). An internal bias analysis of a job evaluation instrument. *Journal of Applied Psychology,* 69, 648-662.

Doverspike, D., Carlisi, A.M., Barrett, G.V., & Alexander, R.A. (1983). Generalizability analysis of a point-method job evaluation instrument. *Journal of Applied Psychology,* 68, 476-483.

Edinberg, M.A., Karoly, P., & Gleser, G.C. (1977). Assessing assertion in the elderly: An application of the behavioral-analytic model of competence. *Journal of Clinical Psychology,* 33, 869-874.

Farrell, A.D., Mariotto, M.J., Conger, A.J., Curran, J.P., & Wallander, J.L. (1979). Self-ratings and judges' ratings of heterosexual social anxiety and skill: A generalizability study. *Journal of Consulting and Clinical Psychology,* 47, 164-175.

Fraser, S.L., Cronshaw, S.F., & Alexander, R.A. (1984). Generalizability analysis of a point method job evaluation instrument: A field study. *Journal of Applied Psychology,* 69, 643-647.

Gillmore, G.M. (1979, March). *An introduction to generalizability theory as a contributor to evaluation research.* Seattle: Washington University, Educational Assessment Center.

Gillmore, G.M. (1983). Generalizability theory: Application to program evaluation. In L.J. Fryans, Jr. (Ed.), *Generalizability theory: Inferences and practical applications.* San Francisco: Jossey-Bass.

Gillmore, G.M., Kane, M.T., & Naccarato, R.W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement*, 15, 1-13.

Gleser, G.C., Green, B.L., & Winget, C.N. (1978). Quantifying interview data on psychic impairment of disaster survivors. *The Journal of Nervous and Mental Diseases*, 166, 209-216.

Hopkins, K.D. (1983). Estimating reliability and generalizability coefficients in two-facet designs. *Journal of Special Education*, 17, 371-375.

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored ratings scales. *Personnel Psychology*, 33, 595-640.

Kane, M.T., Gillmore, G.M., & Crooks, T.J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement*, 13, 171-184.

Kavanagh, M.J., Borman, W.C., Hedge, J.W., & Gould, R.B. (1986, February). *Job performance measurement: A classification scheme for validation research in the military* (AFHRL-TP-85-51, AD-A164 837). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Kavanagh, M.J., MacKinney, A.C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin*, 75, 34-49.

Kraiger, K. (1985, September). *Analysis of relationships among self, supervisory, and peer ratings of performance*. Final report submitted to the AFOSR and AFHRL, Brooks AFB, TX.

Landy, F.J., & Farr, J.L. (1980). Performance rating. *Psychological Bulletin*, 87, 72-107.

Lawler, E.E. (1967). The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology*, 51, 369-381.

Littlefield, J.H., Murrey, A.J., & Garman, R.E. (1977, April). *Assessing the generalizability of clinical rating scales*. Paper presented at the meeting of the American Educational Research Association, New York.

Mariotto, M.J., & Farrell, A.D. (1979). Comparability of the absolute level of ratings of the inpatient multidimensional psychiatric scale within a homogeneous group of raters. *Journal of Consulting and Clinical Psychology*, 47, 59-64.

Morse, D.T., & Morse, L.W. (1976, April). *A model for assessing the effects of departures from reality in performance testing*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Pearlman, K., Schmidt, F.L., & Hunter, J.E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, 65, 373-406.

Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika*, 6, 309-316.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.

Searle, S.R. (1971). Linear models. New York: Wiley.

Shavelson, R.J. (1986, July). *Generalizability of military performance measurements: I. Individual performance*. Paper prepared for the Committee on the Performance of Military Personnel and the Commission on Behavioral and Social Sciences and Education, National Research Council, and National Academy of Sciences, Washington, D.C.

Shavelson, R.J., & Webb, N.M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology*, 34, 133-161.

Smith, P.L (1978). Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Statistics*, 3, 319-346.

Vance, R.J., MacCallum, R.C., Coovert, M.D., & Hedge, J.W. (1988). Construct validity of multiple job performance measures using confirmatory factory analysis. *Journal of Applied Psychology*, 73, 74-80.

Webb, N.M., & Shavelson, R.J. (1981). Multivariate generalizability of General Education Development ratings. Journal of Educational Measurement, 18, 13-22.

Webb, N.M., Shavelson, R.J., Shea, J., & Morello, E. (1981). Generalizability of General Education Development ratings of jobs in the U.S. *Journal of Applied Psychology*, 66, 186-191.

Wieder, G.B., & Weiss, R.L (1980). Generalizability theory and the coding of marital interactions. *Journal of Consulting and Clinical Psychology*, 48, 469-476.